Speed-Accuracy Tradeoffs in Human Speech Production

Adam C. Lammert[1]

MIT Lincoln Laboratory, Lexington, Massachusetts, USA

Christine H. Shadle

Haskins Laboratories, New Haven, Connecticut, USA

Shrikanth S. Narayanan

Signal Analysis and Interpretation Laboratory, Los Angeles, California, USA

Thomas F. Quatieri

MIT Lincoln Laboratory, Lexington, Massachusetts, USA

Classification: Biological Sciences, Psychological and Cognitive Sciences

[1]Corresponding Author: Adam.Lammert@LL.mit.edu

Abstract

Motor actions in speech production are both rapid and highly dexterous, even though speed and accuracy are often thought to conflict. Fitts' law has served as a rigorous formulation of the fundamental speed-accuracy tradeoff in other domains of human motor action, but has not been directly examined in the domain of speech production. The present work seeks evidence for Fitts' law in speech articulation kinematics by analyzing USC-TIMIT, a large database of real-time magnetic resonance imaging data of speech production. A theoretical framework for considering Fitts' law in the domain of speech production is elucidated. Methodological challenges in applying Fitts-style analysis are addressed, including the definition and operational measurement of key variables in real-time MRI data. Results suggest the presence of clear tradeoffs between speed and accuracy for certain types of speech production actions, with wide variability across syllable position, and substantial variability also across subjects. Coda consonant targets immediately following the syllabic nucleus show the strongest evidence of this tradeoff, with correlations as high as 0.72 between movement time and difficulty. Results are discussed with respect to potential limitations of Fitts' law in the context of speech production, as well as the theoretical context. Future improvements in application of Fitts' law are discussed.

Significance Statement: Fundamental tradeoffs between speed and accuracy represents some of the most robust and widely replicated laws of human motor action, having been reported in a wide variety of motor domains. These tradeoffs have not, however, been well-established for speech motor actions, which are some of the most rapid and dexterous that humans execute. Trading relationships between speed and accuracy can provide a window into speech control mechanisms in that they hold promise for explaining specific kinds of speech variability, and also directly relate to prominent functional and neural models of control of directed movement. The present work develops a theoretical basis for speed-accuracy tradeoffs in speech kinematics, and establishes the extent of such tradeoffs in speech through novel analysis.

Keywords: motor control, speed-accuracy trade off, articulatory difficulty, real-time magnetic resonance imaging, Fitts' law, Task Dynamics, VITE

Speed-Accuracy Tradeoffs in Human Speech Production

Introduction

The present work constitutes an effort to apply certain influential ideas of Paul Fitts (Fitts, 1954) to the domain of speech production, specifically his formulation of speed-accuracy tradeoffs in human motor action. Fitts was expressly concerned with quantifying the capacity of the human motor system to perform motor actions. One important outcome of that work was a rigorous formulation of perhaps the most robust and widely replicated laws of human motor action. It was shown that for discrete, targeted actions, the time taken to complete a movement displays a linear relationship with task difficulty, where difficulty is a function of movement distance and the tolerable error in reaching the target. This now well-known relationship has subsequently been referred to as *Fitts' law*, and has been used widely to model speed-accuracy tradeoffs in a variety of human movement domains. Example application domains include manual pointing and reaching (as in Fitts' original study), targeted foot movements (Drury, 1975), balance and posture (Duarte & Freitas, 2005), and computer device interaction (Card et al., 1978). Fitts' law has also been applied to ballistic movements, including eye saccades (Ware & Mikaelian, 1987), although there is meaningful debate over whether movements that do not rely heavily on feedback are subject to the same law (Carpenter, 1988; Sibert & Jacob, 2000; Drewes, 2013).

It is not well-established whether this pervasive law of human movement is obeyed by speech motor actions. Despite evidence that speech articulation obeys related tradeoffs among metrics of speed, distance and curvature (Lofqvist & Gracco, 1997; Perrier & Fuchs, 2008; Kato et al., 2009), Fitts' law has not been directly examined in the context of speech production. Motor actions associated with speech production are some of the most rapid and dexterous that humans execute. The presence of speed-accuracy tradeoffs would imply, however, that it is not necessarily possible to attain high levels of speed and accuracy at the same time. Moreover, in speech production, there are potentially multiple domains in which accuracy may be demanded, ranging from articulatory and acoustic, to prosodic and

communicative, with all of these demands being possibly simultaneous and overlapping. Kinematics are the present focus because many human motor actions exhibit a clear kinematic tradeoff between speed and accuracy. The present paper examines one aspect of accuracy in speech actions: the kinematics of "reaching" for maximal articulatory targets. Articulatory speech actions can be conceptualized as discrete motor actions

Speed-accuracy tradeoffs can provide a window into the control mechanisms of directed movements. While it is possible that biomechanical constraints exist that give rise to such tradeoffs, there is also good reason to believe that they are the result of properties of planning and control. It can be shown that Fitts' law is consistent with traditional models of feedback-driven motor control (Langolf et al., 1976). Moreover, it is closely related to models of neural dynamics of movement trajectory formation (Bullock & Grossberg, 1988). If it is true that control mechanisms bring about speed-accuracy tradeoffs, it implies that changes in timing can be used to assess demands in accuracy and, conversely, that changes in accuracy can be partially attributed to speaking rate demand.

The presence of Fitts-type tradeoffs in speech production would help to explain a variety of observed phenomena. It has been argued that speech motor actions vary considerably in difficulty, and that differences in difficulty relate to elements of timing. Hardcastle (1976) asserted that the difficulty (or complexity, to use his terminology) of an articulatory action should be defined in terms of both the number of articulatory variables that are recruited over the course of that action, and in terms of the precision required for each of those variables. The issue of articulatory precision and its kinematic consequences is entirely compatible with Fitts' law. Hardcastle goes on to make direct reference to a speed-accuracy tradeoff in speech production, while arguing that fricatives require more precision than stop consonants: "One of the possible effects of this greater precision is that the articulators involved in the production of a fricative might move more slowly than for the production of a stop." Hardcastle notes that this may help to explain why vowels are often lengthened in advance of fricatives (i.e., more time is required to execute the more

difficult fricative articulation) – as originally suggested by MacNeilage (MacNeilage, 1972) – and lower vowels are longer than higher vowels (Lehiste, 1970) (i.e., more time is required for the tongue to travel the longer distance). This is also a possible explanation for the observation that fricatives have longer durations, in general, than stops (Kuwabara, 1996).

Speed-accuracy tradeoffs may also aid in explaining changes in speed and accuracy during speech acquisition and loss. Speaking rate decline is associated with various kinds of neurological decline (Yunusova et al., 2008; Williamson et al., 2015). Change in speaking rates are perhaps a compensatory mechanism in order to maintain accuracy when difficulty increases. Even in normal speakers, accuracy and intelligibility decline at markedly increased speaking rates (Kleinow et al., 2001; Krause & Braida, 2002). The notion of articulatory difficulty may also help to explain why fricatives tend to be acquired later than stops (Templin, 1957), and why some productions are more quickly impacted when the condition of the motor system changes, as in the idea that sleepiness and alcohol intoxication lead to the salient changes in fricatives associated with "slurred speech" (Chin & Pisoni, 1997; Schuller et al., 2014). Better knowledge of the presence and nature of speed-accuracy tradeoffs in speech, therefore, would have natural applications toward identifying those elements of speech production that are early indicators of neurological change or decline.

The purpose of this paper is three-fold. The primary goal is to analyze speech articulation using a large database of real-time magnetic resonance (rtMRI) data, in order to assess whether articulatory kinematics conform to Fitts' law. A second, associated goal is to address the methodological challenges inherent in performing Fitts-style analysis on rtMRI data of speech production. Methodological challenges include segmenting continuous speech into specific motor tasks, defining key variables of Fitts' law in the domain of speech articulation, and deciding how to operationalize these definitions and extract related measures from complex and high-dimensional rtMRI data. Finally, a third goal is to present a novel mathematical argument for Fitts' law in speech production, and make a

theoretical argument for why one would expect to observe behavior consistent with the law. Section 2 gives a brief introduction to the concepts and mathematics behind Fitts' law, and presents an argument for Fitts' law in speech production. Section 3 describes the data used in the present study, and the necessary pre-processing for the task being considered. Section 4 explains the present approach to applying Fitts' law in the domain of speech production data. The results of applying the proposed methodology to rtMRI data, and a discussion of the results in terms of the goals of the paper, are given in Section 5. Lammert et al. (2016) have previously reported on an initial effort to meet some of the goals of the present work by analyzing portions of the USC-TIMIT database and forming necessary elements of the data analysis. This paper constitutes a substantial expansion of that work, providing a more extensive and deeper analysis on more subjects, as well as a better developed framework for considering speed-accuracy trade offs in a speech production context. In particular, the present work provides (1) an analysis of six additional subjects, altogether comprising the entirety of the real-time data from the USC-TIMIT database, (2) a more detailed look at speed-accuracy relationships in speech tasks of different varieties, specifically tasks situated in different parts of the syllable, and (3) elucidation of a theoretical framework for considering Fitts' law in the domain of speech production, and its mathematical connection to prominent models of speech motor control and neural control of movement.

Background

Fitts' law can be stated precisely in mathematical terms. It has deep connections with several prominent frameworks of directed human motor control. This section is intended to provide an overview of Fitts' law, including the mathematical statement thereof, as well as connections to the Task Dynamics control framework Saltzman & Kelso (1987); Saltzman & Munhall (1989), and the VITE model of neural control of directed human movement (Bullock & Grossberg, 1988).

Statement of Fitts' Law

Given a *target* associated with a given task, as well as an *initial position* (also, *context*), key parameters of that action can be defined, and incorporated into a simple framework that represents the difficulty associated with that task. One parameter is the *distance* to the target from the initial position. Longer distances are assumed to make a task more difficult. The other parameter is the *width* of the target, which represents the tolerable error in reaching the target. A wider target is assumed to make a task less difficult, perhaps corresponding to more slack being permitted in declaring an action successful. The ratio of the distance to the target, D, and its width, W, are then associated with the *index of difficulty* (ID) in the following way:

$$ID = \log_2\left(\frac{2D}{W}\right) \tag{1}$$

The ratio D/W constitutes one definition of the precision of a task. Taking the base-2 logarithm of this precision, then, gives the ID units that can be interpreted as bits, inspired by Claude Shannon's information theory (Shannon & Weaver, 1949). The ID, having encapsulated a notion of precision of action, should then be related to the *movement time* (MT) associated with a given task, under the hypothesis that a tradeoff exists between speed and accuracy of that task. This relationship, Fitts' law, is commonly formulated as a simple, linear one:

$$MT = a \cdot ID + b, \tag{2}$$

where a and b are constants, the values of which depend on the task and characteristics of control. Fitts' law has been derived in various ways since the original formulation (Crossman & Goodeve, 1983; Bullock & Grossberg, 1988; Beamish et al., 2006).
Note that, whereas the distance associated with a task is typically fairly straightforward to define given an initial position and a target (e.g., the Euclidean distance), the width parameter has been defined in many different ways. Fitts' original experiments included targets with a literal, physical width of varying size, but many experimental setups have

only a point target (as assumed in many human actions). In the domain of speech production, however, one is faced with an added complication stemming from a lack of consensus regarding how an articulatory target should be defined, or indeed whether an *articulatory* target (as opposed to acoustic) exists at all. In the present work, it is assumed that articulatory targets do exist, following the specific definition explained below.

It is worth noting certain subtleties with regard to the interpretation of Fitts' law as an expression of a speed-accuracy trade off. Much of the literature related to Fitts' law interprets the law as such a trade off, either implicitly or explicitly. For example, Fitts & Radford (1966) discuss the variables MT and W as representing speed and the reciprocal of accuracy. The law, under this interpretation of the variables, is therefore an expression of a speed-accuracy trade off, with that additional caveat that accuracy must always be considered relative to D. This interpretation of Fitts' law assumes that "speed" is the reciprocal of MT – essentially an expression of the speed of completion of the task – rather than articulator speed, as in the classical-mechanical sense of $|D/MT|$. A classical definition of the speed-accuracy trade off might be $W_1 = cD/MT$, stating that W_1 is proportional to articulatory velocity, given some coefficient c. This classical definition is not exactly the same as Fitts' law, but the two can be related be rewriting Equation 2 2 as: $MT = \log_2(cD) - \log_2(W_2)$, implying that $W_2 = cD/2^{MT}$ (ignoring coefficients, for simplicity, and substituting c for the value 2). The quantity $cD/2^{MT}$ is still not the classical definition of speed, but it similarly decreases monotonically with MT, and the quantities W_1 and W_2 from the Fitts' and classical definitions can be related by a multiplicative factor, $W_1 = \eta W_2$, where $\eta = 2^{MT}/MT$.

Theoretical Framework

Fitts' law has substantial mathematical connections with the dynamical systems view of coordination and control of human movement (e.g., Turvey (1990); Davids et al. (2003)). This section attempts to elucidate those connections, and to provide a novel argument for

expecting behavior consistent with Fitts' law in speech production on the basis of prominent theories of speech motor control and neural dynamics. Within the dynamical systems perspective, one representative body of work that has had an impact on modeling and explaining speech articulation is that of Task Dynamics Saltzman & Kelso (1987); Saltzman & Munhall (1989). Task Dynamics constitutes a control system that allows for the description and achievement of directed actions in a relatively high-level *task space*, as opposed to the relatively low-level *articulator space*, defined by variables of mobility such as muscle activations. An example of a task space for a manual reaching task would be three-dimensional Cartesian space, as opposed to the articulatory space of joint angles at the shoulder, elbow and wrist. Task space for a speech production action could be the space defined by the first three formant frequencies, or the space defined by vocal tract constriction degree and location, as in Articulatory Phonology (Browman & Goldstein, 1992). These high-level spaces are the natural spaces in which to define the goals of directed action, and Task Dynamics defines a rigorous framework in which motor commands can be generated in articulator space toward the completion of movements in task space.

In Task Dynamics, the targets of directed movement are assumed to be points in task space. Those targets are achieved by point-attractor dynamics, governed by 2^{nd}-order equations of motion consistent with a critically damped harmonic oscillator. the dynamics of which are well understood from classical mechanics. For the sake of simplicity, consider a one-dimensional task space. The equations can be written as follows:

$$\ddot{X} = \frac{-c\dot{X}}{m} - \frac{k(X - X_0)}{m}, \tag{3}$$

where X is the displacement of the controlled variable and X_0 is the target. The forward dynamics take the form of a second-order dynamical system, conforming to Equation 3, that transforms the error signal, ΔX, into the second derivative of the articulator-space variable u. An overview of the control flow in Task Dynamics is shown in Figure 1. Equation 3 is contained within the box labelled "Forward Dynamics", which computes the acceleration of u from $\Delta X = X_0 - X$. Note that the low-level articulator variables, u, and

the relevant kinematic transformations between task and articulator spaces are not discussed in the present context. This is because dynamics in task space only are sufficient to account for Fitts' law.

Fitts' law can be seen as a direct consequence of such dynamics. A mathematical connection can be made through an examination of the step response of the system, which corresponds to the sudden appearance of a new target in task space. The relevant quantity then becomes the settling time of the damped harmonic oscillator, that is, the time required for the system to converge within a certain percentage of the final target value, beginning at rest. It is well known from classical mechanics that, in the case of critical damping, the rate of convergence in the step response to a change in target follows a decaying exponential. That is, the displacement of the system at time t is $X_t = X_0 e^{-\omega_0 \zeta t}$, in a system where the natural frequency is $\omega_0 = \sqrt{k/m}$, and the damping ratio is $\zeta = c/2m\omega_0$. In the case of critical damping, $\zeta = 1$, and $X_t = X_0 e^{t\sqrt{k/m}}$.

Several of these quantities can be related directly to those in the formulation of Fitts' law. The value t can be considered as MT, the time at which the system is considered to have settled, or completed its action. Given that the movement takes time t to complete, and X_t is the residual displacement of the controlled variable after the action has completed, X_t can be equated with the error tolerance W. Furthermore, X_0 is equivalent to the movement distance, D, if the movement is considered to begin at $X = 0$. Following from these identities, we can express the step response equation above with a change of variables, as $W = De^{-\sqrt{k/m}MT}$. This can be easily rewritten as:

$$MT = \frac{1}{\sqrt{\frac{k}{m}}} \cdot \ln\left(\frac{D}{W}\right), \tag{4}$$

which is already similar to Fitts' law in form. We can find the conditions under which they are equivalent by setting this new formula for MT equal to the one taken from Fitts' law. Beginning – for the sake of clarity – with a change of logarithm base from the law expressed

in Equation 2 (corresponding to a switch of units in ID from *bits* to *nats*), we have:

$$a \cdot \ln\left(\frac{2D}{W}\right) + b = \frac{1}{\sqrt{\frac{k}{m}}} \ln\left(\frac{D}{W}\right) \tag{5}$$

It is easy to show that this equation holds for certain values of a and b. For instance, assuming that $a = \frac{1}{\sqrt{k/m}}$ (the reciprocal of the natural frequency of oscillation), one can solve to find that $b = \frac{\ln(2)}{\sqrt{k/m}}$. Therefore, Fitts' law conforms to the predicted kinematic behavior of a damped harmonic oscillator, which is consistent with the behavior of a Task Dynamic control system when acting to achieve a specific movement target.

In addition to the kinematic considerations of the Task Dynamics model, Fitts' law also has substantial mathematical connections with models of the neural dynamics underlying the dynamical systems view of human motor control. An influential neural-inspired network model for explaining kinematic trajectory formation of directed movement is the VITE model (Bullock & Grossberg, 1988). This model's predictions are highly consistent with those of the Task Dynamics model, owing to the fact that VITE is a 2^{nd}-order dynamical system much like Task Dynamics (as pointed out by, e.g., Beamish et al. (2006)). VITE comprises a network of interacting hypothesized neural populations which generate a movement command, given some target position. The neural populations are configured in order to code distinct quantities that are needed in the generation of the motor command. Among the interacting neural populations, there is (a) a population representing the target position command (TPC), (b) a population representing the present position command (PPC), and (c) a population referred to as the difference vector (DV) population, which represents the difference between the PPC and TPC.

The specific structure of VITE's interacting network is shown in Figure 2. Note the many similarities of this structure to that of Task Dynamics in Figure 1. TPC, as a representation of the target position, produces a target position X_0. The DV population compares the target to the system's current position, and computes the task-space dynamics of the network. The PPC population, meanwhile, integrates the DV population activation into position information, in analogy to the physical plant in the Task Dynamics

control flow. The network dynamics have the following form:

$$\dot{V} = \alpha(X_0 - X - V), \tag{6}$$

and

$$\dot{X} = GV \tag{7}$$

where the parameter α has been termed the "convergence coefficient" and G is the "go" signal, which initiates and sustains movement. These equations also compare easily to the equations of motion for Task Dynamics given above in Equation 3. There are important differences, however [2]. First, all computations are done at the level of tasks, with no mention of the articulator space. Therefore, there is no need for kinematic transformations between task space and articulator space in VITE. Second, the inclusion of G has no equivalent in Task Dynamics, where it is assumed (implicitly) that movement toward a target is always active as long as the target exists.

As with Task Dynamics, Fitts' law can be seen as a direct consequence of these neural-inspired dynamics. This can be shown by demonstrating the mathematical relationship between the equations of motion in Equation 6 and Equation 3. If $G = 1$, then $V = \dot{X}$, and subsequently that 6 and 7 collapse into the single equation:

$$\ddot{X} = \alpha(X_0 - X - \dot{X}), \tag{8}$$

which is the same as Equation 3, if $\alpha = -k/m = -c/m$. Therefore, VITE is consistent with Task Dynamics control, and Fitts' law can be seen as related to both those models in a general sense, and as a direct consequence of them under the specified conditions and parameters. Note, incidentally, that because the damping coefficient in VITE is fixed at $c = -\alpha m$, in order for the system to be critically damped (i.e., $c = 2\sqrt{mk}$), as in Task Dynamics, that $m = k/4$. Overdamping will occur with $m > k/4$, and underdamping with $m < k/4$.

[2] Also, this presentation glosses over a nonlinearity in the original VITE formulation, where V is not allowed to go negative. This detail was not seen as important in the present context.

Practical Framework

To apply Fitts-style analysis to speech production data, it is necessary to operationally define the targets of articulation in space and time. To that end, it is assumed that a single articulatory target is associated with each phoneme. Targets might not be reached during continuous speech for a variety of reasons, including undershoot, misarticulation, or tolerance of the controller to some deviation from the target. However, it is assumed that the action associated with a given phone comes closest to achieving its target at the temporal center of the associated phone interval. Thus, each targeted task in continuous speech can be conceptualized as movement from one phoneme target to another, constituting a specific diphone. Tasks conceptualized this way can also be referred to by diphone, which represents a context-target task pair. It is further assumed that the target of a given phoneme is a vector in high-dimensional articulatory space. The location of that vector is estimated as the mean of all tokens with a given phoneme label. The initial position for a given task is assumed to be the target immediately preceding the current one. All these notions will be defined formally below and in Figure 3.

It has been well-established that the temporal relationship between speech gestures varies as a function of their positions within the syllable (Browman & Goldstein, 1995; Krakow, 1999; Byrd et al., 2009). Therefore, it was hypothesized that adherence to Fitts' law might vary depending on the task type, where type was determined by syllable position. To facilitate analysis of speech tasks conditioned on syllable position, a syllabification was performed five categories of interest were defined with respect to syllable structure (see Figure 6):

- Category 1: Onset-Nucleus Task (initial position: final onset consonant; target: syllable nucleus)
- Category 2: Nucleus-Coda Task (initial position: syllable nucleus; target: first coda consonant)

- Category 3: Onset-Onset Task (initial position: onset consonant; target: succeeding onset consonant)
- Category 4: Coda-Coda Task (initial position: coda consonant; target: succeeding coda consonant)
- Category 5: Coda-Onset Task (initial position: final coda consonant; target: first onset consonant of succeeding syllable)

Note that the tasks in category 5 are across syllables, whereas the tasks in categories 1–4 are all within a single syllable.

Method

Data, Pre-Processing & Feature Extraction

Data used in the present study are from the USC-TIMIT database (Narayanan et al., 2014). USC-TIMIT is a publicly-available collection of speech production data from speakers of American English. Speech articulation data were gathered for the database using two different modalities, rtMRI and electromagnetic articulography (EMA). The rtMRI data were used in the present analysis. Resolution of the rtMRI data is 68 by 68 pixels, with pixels 2.9 by 2.9 mm in size, at a frame rate of 23.18 frames/s. Audio was simultaneously recorded at a sampling frequency of 20 kHz, and later subjected to noise cancellation (Bresch et al., 2006).

The rtMRI data from all five male (M) and five female (F) subjects from the database (i.e., M1-5 and F1-5) were used in the present analysis. Forced phoneme alignment was carried out using SAIL-Align (Katsamanis et al., 2011). Subjects were analyzed separately, due to concerns about the proper method of combining articulatory features across subjects. Subjects read aloud the 460 sentences constituting the MOCHA-TIMIT corpus (Wrench, 1999). For three of the speakers, software-related difficulties resulted in MRI frames going unrecorded in the data, which makes ideal audio-video synchronization impossible.

Sentences in which this problem arose were discarded. In the end, 346 of the 4600 total sentences were discarded, including 175 from F4 (sentences 286 to 460), 166 from M5 (sentences 295-460) and only five sentences from M3 (sentences 331-335). All 460 sentences were represented in the data for the other seven subjects.

The analysis presented here began by treating the gray-scale intensity values of each pixel in the image plane as a candidate articulatory feature (Lammert et al., 2010; Lammert, Ramanarayanan, et al., 2013). These candidate features were pre-processed and recombined prior to analysis, in order to produce new features that are fewer in number and more specific to speech articulation (details below). Such a pixel-wise approach may seem unintuitive, but it provides the opportunity to analyze data about the entire midsagittal plane, while making minimal assumptions about what information might be important for describing articulation. Pixel-wise analysis is also relatively robust compared to a more traditional edge-detection and boundaries-extraction approach when applied to low-contrast, low spatial-resolution rtMR images (Lammert et al., 2014).

The rtMRI image sequences were pre-processed to facilitate further analysis, in particular to (a) isolate frames of interest, and (b) reduce the high dimensionality of the data to a manageable number. Analysis began with an image sequence, X, of the form $X = [I_1 I_2 I_3 \dots I_n]^T$, comprising all n image frames I_m in the corpus from a single subject, where the images I_m are vectorized in column format. That is, pixels located at (i, j) in rectangular r by c image format are now located at $c(i-1)+j$ in the vector I, and I is of length rc. Prior to further analysis, images underwent an intensity correction procedure to compensate for the reduction in coil sensitivity moving posteriorly, from the lips toward the pharynx (i.e., at increasing spatial distance from the coil). A retrospective correction scheme was implemented, incorporating a nonparametric, monotonically increasing estimate of coil sensitivity, which was derived from all pixel values in the video sequence (Lammert, Ramanarayanan, et al., 2013). Decreasing coil sensitivity results in lower mean and smaller dynamic range of intensity values for pixels at large distances from the coil.

Intensity correction must be done to ensure that pixel intensity values can be compared and interpreted across all spatial locations. Image intensity correction results in a matrix X^c of corrected image vectors.

Pixels that are unrelated to vocal tract action were eliminated by a simple threshold procedure. Pixels representing the air around the head, or representing static spinal or brain tissue, have intensities that change very little over the image sequence. These pixels can be identified by calculating the variance along columns of X^c, and selecting only columns with highest variance. Such pixels represent approximately 75% of all pixels in the images analyzed in the present work, as identified by visual inspection of the images. Therefore, the matrix X^c_{sub} was formed, which contained only those columns of X^c with variance above the 74^{th} percentile across all columns.

The matrix X^c_{sub} is therefore n by $rc/4$ in size, but only a subset of the n data vectors represent vocal tract configurations temporally close to an articulatory target. Using the above operational definition of articulatory targets, the row vectors in X^c_{sub} corresponding to the temporal centers of phones are identified and extracted. From the forced alignment, each phone is assigned a starting boundary A_m, and an ending boundary B_m, both in seconds. From these, the temporal center of a phone can be calculated as $\Gamma_m = (A_m + B_m)/2$, and the corresponding image frame is $\arg_m \min(\Gamma_m - \tau_m)^2$ for timestamps $\tau_1, \ldots, \tau_n$ associated with each original image frame. In this way, a new matrix Y is formed, which is P by $rc/4$ in size, where P is the total number of phones represented in the image sequence, $P \approx 15121$.

Principal Component Analysis (PCA) was employed to further reduce the data dimensionality. $Z = YC_L$ was computed, where C is the matrix whose columns are eigenvectors of YY^T, and C_L is a matrix containing only L columns that represent eigenvectors with the highest eigenvalues (i.e., the largest principal components). The magnitude of L was chosen so as to retain $\geq 85\%$ of the variance for each subject being analyzed. Across the subjects analyzed in the present study, L was approximately equal to

50. The resulting P by L matrix Z, which contains a reduced-dimension representation of each vocal tract configuration nearest to an articulatory target, was used for all subsequent analyses. Images illustrating the key stages in this image pre-processing pipeline are shown in Figure 4. It is worth noting that PCA performed with the correlation matrix, rather than the covariance matrix, might provide an alternative method of dimensionality reduction that would also obviate the need for image intensity correction.

Syllabification was performed based on the forced alignment results, beginning with the word-level transcription from the adaptive forced alignment procedure. Words were translated into phoneme sequences finding their entries in the CMU Pronouncing Dictionary, which are already syllabified. Syllables were subsequently divided into onset, nucleus and coda by identifying the vowel as the nucleus, and considering all phones preceding the nucleus as part of the onset, and all phones following the nucleus as part of the coda. This syllabification allowed for (a) partitioning tasks into the meaningful categories of interest with respect to syllable structure, and (b) calculation of syllable position-specific movement times.

Distance & Width Calculations

For the purposes of analysis, a phoneme vector Π is defined, which is of length P. The p^{th} element of Π, Π^p, is a numerical index from 1 to 35, uniquely specifying an American English phoneme, and representing the phoneme associated with row p of Z. The vector S_g, which is of length P, is associated with a given phoneme index g from 1 to 35. $S_g^p = 1$ whenever $\Pi^p = g$, and 0 elsewhere. The mean configuration vector associated with the phoneme indexed by g is

$$F^g = \frac{\mathbf{1}^T diag(S_g) Z}{\parallel S_g \parallel_1} \tag{9}$$

where $\mathbf{1}$ is a vector of ones. The vector F^g represents our operationally-defined articulatory target associated with the phoneme indexed by g.

For every pair of phoneme indices g and h, it is now possible to state precisely the spatial

distance between the associated phonemes. Using the Euclidean distance in the L-dimensional articulatory space, the distance $D_{gh} =\| F_g - F_h \|$. A graphical representation of this can be seen in Figure 5.

To calculate the time to reach phoneme h from g (indices), assume that S_{gh} is a vector that is 1 whenever both $\Pi_p = h$ and $\Pi_{p-1} = g$. Similarly, S_{hg} is 1 whenever both $\Pi_p = g$ and $\Pi_{p+1} = h$. The mean time, then, between the phonemes indexed by g and h across all instances is

$$T_{gh} = \frac{\mathbf{1}^T diag(S_{gh})\Gamma - \mathbf{1}^T diag(S_{hg})\Gamma}{\| S_g \|_1} \tag{10}$$

As mentioned in the discussion of syllabification above, it was hypothesized that adherence to Fitts' law might vary depending on syllable position, because the temporal relationships between speech gestures are known to vary as a function of syllable position. Therefore, T_{gh} was, in fact, calculated also for each of the five syllable position-based categories listed above. When the linear relationship between ID and T was assessed on the category-by-category basis, it was this category-specific value for T_{gh} that was used.

There are many possible definitions for the width of the target in a speech production task. There are no hard physical limits around the target, as in Fitts' original experiments, which necessitates exploring other definitions. Width could be defined in terms of variability about the target, as in later measures of "effective" width (Welford, 1968; Fitts & Peterson, 1964). Other definitions have been based on the amount of under/overshoot associated with a particular movement (Bullock & Grossberg, 1988). However, the nature of speech being such that phonetic contrasts can be made with very small changes in vocal tract configuration, allows for the possibility of another definition based on the density of targets in articulatory space. Consider the distance values D_{fh} for a given h and all $f = 1, \ldots, 35$. These distance values with respect to h can be sorted and ranked, and – given a parameter k – we can select the distance between F_h and the k^{th} closest vector F_{f_k}. That distance can be used as the basis for a high-dimensional k-nearest-neighbor density calculation. The

probability density of configuration vectors in the neighborhood of F_h will be:

$$Q_h = \frac{k}{35\frac{\pi^{L/2}}{\Gamma(\frac{L}{2}+1)}D_{f_k h}^L} \tag{11}$$

where $\Gamma(x)$ is the gamma function and 35 is the number of phonemes under consideration (24 consonants and 11 vowels, with no diphthongs or rhoticized vowels). The width can be calculated from this probability density as $W_g = -\log_2(Q_g)$. Note that the final width value does not depend on the context.

Fitts' law can be calculated directly using D_{gh}, T_{gh} and W_h for any phoneme indexed by h, and presented in the context of another phoneme g. Applying Equation 1, it is possible to calculate $ID_{gh} = \log_2(2D_{gh}/W_h)$. Furthermore, by Equation 2, we expect that $T_{gh} = a \cdot ID_{gh} + b$, for some coefficients a and b. Images of initial positions and targets for one example each of high- and low-ID tasks are shown in Figure 7.

Results and Discussion

The strength of the relationship between MT and ID was assessed using linear correlation (Pearson's r), in keeping with the linear form of Fitts' law. The correlation coefficients are shown in Table 1, divided by syllable position-specific category and by subject. Performing this analysis separately for each subject, and once for each task category, means that a total of $5 \times 10 = 50$ individual correlations were calculated, with 50 corresponding tests for statistical significance. It therefore became necessary to consider the significance of these results in light of some kind of multiple comparisons adjustment. It is not clear whether or how much these individual correlations are dependent, so statistical significance of the correlation coefficients is shown at three distinct threshold values: $\alpha = 0.05$ (Fisher's traditional value), $\alpha = 0.01$ (an intermediate value) and $\alpha = 0.001$, which is the conservative, Bonferroni-adjusted threshold value.

Subject M2 had the generally highest correlation values of all subjects, indicating that the Fitts-style relationships were strongest and clearest for that subject. Figure 8 shows MT versus ID for subject M2, showing the strength and nature of those relationships for each of

the syllable position-specific task categories. The correlation values corresponding to each category, and the associated p-values, are shown above each plot.

Articulatory Difficulty

Results suggest that the difficulty associated with targeted articulatory kinematics is highly variable in speech production. ID ranges from approximately 0.25 to 1.75 bits for all subjects. A few general patterns in the distribution of ID can be noted. Difficulty was assessed by looking at the overall average ID associated with a given target position. ID values for each subject were normalized between 0 and 1, in advance of taking the mean ID for each task across subjects. The mean ID was then calculated for each task with a consonant target, given a vowel initial position. These tasks, listed from most difficult to least difficult, were: ʒ, ʧ, θ, h, ʃ, p, w, ʤ, ð, b, g, k, j, f, ŋ, v, z, s, m, l, d, r, n, t. The mean ID was also calculated for each task with a vowel target and a consonant initial position. These tasks, listed from most difficult to least difficult: ʊ, o, ɑ, ɔ, e, æ, u, ɛ, i, ɪ, ə. Figure 4 shows example low- and high-ID tasks for subject M2.

Consonant tasks involving labial articulation, whether primary (/p/, /w/, /b/, or secondary /ʃ/), tend to have a higher difficulty. Nasals and liquids all ranked as lower difficulty. Fricatives /ʃ/, /θ/, /ð/ and affricates show higher ID, a fact that is consistent with Hardcastle's assertion that fricatives – and perhaps by extension, affricates – are the sounds of speech requiring the greatest accuracy. This does not necessarily include all fricatives, however, as /z/ and /s/ were ranked relatively lower. The difficulty associated with producing fricative and affricates is particularly evident when examining in the context of low, back vowels, where the distance from the initial position to the target position is lengthened. One can also observe that these articulatory tasks require more time to complete than other tasks. Also consistent with Hardcastle is the observation that stop consonants – particularly alveolar – require little accuracy, and are therefore not difficult. Since distance is a factor under consideration, one can see that this effect is again

emphasized when the initial position is a high, front lax vowel. It is important to remember that sibilants may have complex aerodynamic requirements that will not necessarily show up in the purely kinematic analysis in the present work.

Vowel targets that were low and back had a higher level of difficulty, as compared to the relatively lower difficulty high and front vowels. Vowels that were not directly along this primary low-back/high-front axis, including the high-back vowel /u/ and the low-front vowel /æ/ appeared, together toward the middle of the vowel difficulty ranking. Schwa was ranked as the least difficult vowel to produce. This ranking is consistent with the importance of D in computing ID. With a schwa target, the speech articulators should have, on average, a shorter distance to travel from other initial positions. It has been shown that, although it has a distinct phonetic identity, schwa is perceptually and articulatory similar to "articulatory setting", which is the neutral posture from which speech actions are deployed and to which they tend to return (Ramanarayanan et al., 2013). This neutral posture is hypothesized to be kinematically advantageous, just as there is evidence that it is mechanically advantageous (Ramanarayanan et al., 2014). Conversely, low-back vowels should require the speech articulators to travel longer distances from a variety of initial positions, in order to reach kinematic targets in the region of the pharynx.

One other issue of note concerns that fact that the distance and width parameters do not seem to contribute equally to the index of difficulty, given the present definitions, and current data. Although both parameters influence the final value of ID, difficulty seems to be determined to a much larger degree by distance than by width. For instance, for subject M2, the correlation between D and ID across all diphones is substantially greater (Spearman's $\rho = 0.987$, $n = 1190$, $p \ll 0$) than the correlation between W and ID (Spearman's $\rho = 0.222$, $n = 1190$, $p \ll 0$). Similar trends are seen across all subjects. Note that this correlation between W and ID is in the opposite direction from expected, based on the equation for ID. This may be due to the fact that, for these data, D and W seem to be positively correlated (e.g., for subject M2: Spearman's $\rho = 0.3565$, $n = 1190$, $p \ll 0$).

Speed-Accuracy Tradeoff

Results suggest that targeted speech actions exhibit a clear tradeoff between speed and accuracy in certain task categories, and with substantial interspeaker variability. Significant correlations can be seen in the data that correspond to the relationship between MT and ID predicted by Fitts' law. The strength of that relationship varies across speaker and task type. The strongest and most highly significant of such relationships are seen for Nucleus-Coda tasks across all subjects. Onset-Nucleus and Coda-Onset tasks also showed generally high correlations that were significant for at least three subjects (M2, F1 and F2, but also M4 for Coda-Onset tasks). Note that many fewer Onset-Onset and Coda-Coda tasks exist, as compared to other task types. For speakers that display significant correlations between ID and MT, the relationships appear linear (see, e.g., Figure 8), but with abundant added noise around the trend line. There is also a notable deviation from linear at small ID values, where MT appears to hit a minimum value around 50ms. This floor effect may reflect physiological constraints on the production apparatus.

Despite several significant correlation values between MT and ID, the correlations observed in the present analysis are relatively modest compared to those observed in other domains of human movement. Correlation coefficients above 0.9 are commonly reported in the literature (MacKenzie, 1992), whereas the best correlation value observed in the present analysis was 0.72 (Subject M2, Nucleus-Coda task). The correlation values are also highly dependent on the task type and the subject under consideration. One question raised by such results is why this seemingly fundamental tradeoff, that has been well-established in other motor domains, appears to be somewhat weakly and variously obeyed in the domain of speech production. There are several potential explanations, which are considered below. The class of movements considered ballistic (i.e., occurring without feedback control while movement is underway) provide an example explanation to consider. It has been argued that ballistic movements, including eye saccades, do not obey Fitts' law because a lack of feedback means that movement time does not depend on the required accuracy of the task,

but only on the movement amplitude (Carpenter, 1988; Drewes, 2013). It is possible that certain speech production actions implement ballistic control. However, even despite their rapidity, speech motor tasks are typically not modeled as being ballistic in nature. Major models of speech motor control involve feedback at fine temporal scales. It has already been discussed that the Task Dynamics model relies on feedback, and leads naturally to precisely the kind of speed-accuracy trade offs described by Fitts law (Saltzman & Munhall, 1989). Other prominent models of speech motor control, such as the DIVA model (Guenther et al., 1998) and State Feedback Control (Houde & Nagarajan, 2011), also rely on feedback, although the connection to Fitts' law has not been explicitly made. Therefore, rather than concluding that each of these models is inaccurate, and that ballistic control of speech movements provides a better explanation of the present data, a more likely explanation for the weak and variable observed relationships is that the definition of speech tasks used in the present work needs to be revised in one of several ways.

One way to reconsider the presently-used definition of speech tasks is to make them multimodal, for instance by incorporating prosodic constraints. As mentioned above, speech has multiple levels in which accuracy may be demanded. Speech motor actions have communicative and prosodic goals, in addition to kinematic requirements. Temporal constraints exist as part of those goals, both at the level of phonetic segments (e.g., lengthening as a phonemic contrast) and suprasegmentally (e.g. accenting). Indeed, the objective function for speech motor control might be formulated as an information-theoretic measure exemplifying both the achievement of the kinematic goal and any temporal information encoding, including active and incidental temporal aspects. A modification of speech tasks (and, perhaps, Fitt's law itself) is needed to account for these various levels of task requirements, and associated timing requirements.

Another important change to the presently-used definition of speech tasks may be to account for non-sequential, overlapping articulatory targets, as opposed to the purely sequential tasks considered in this work. In fact, the results already indicate the need for

such an enhancement. It has been well established that speech articulatory gestures at certain positions in the syllable are highly overlapping, whereas others are more sequential (Browman & Goldstein, 1995; Krakow, 1999; Byrd et al., 2009). Specifically, onset consonants would be expected to overlap with each other extensively, and would overlap with the succeeding nucleus, as well. The nucleus, in contrast, should overlap very little with the succeeding consonant in the coda. The present results clearly show that the correlations are strongest for the Nucleus-Coda tasks across all subjects, which is exactly what would be predicted by the sequential, non-overlapping nature of the gestures involved at that position in the syllable. The Onset-Onset tasks, for which the assumption of sequential targets may be inappropriate, show the poorest overall correlations. Moreover, speech tasks should potentially also allow for contextually modified targets. Enhancing speech tasks in this way would also allow for a natural way to capture co-articulation in targets, as opposed to the fixed targets considered in this work.

There are substantial interspeaker differences in the strength of correlations between ID and MT. These differences are evident in the Nucleus-Coda tasks, where most subjects displayed significant correlations, but to different degrees. Interspeaker differences are also evident for other tasks, such as the Onset-Nucleus tasks, where some subjects showed marginal correlations (e.g., M3 and F5) and others (e.g., M2 and F2) showed highly significant correlations. Important questions remain regarding an explanation for this prevalent interspeaker variability. These differences may reflect interspeaker differences in control strategies, that in turn are a function of speaking rate, age, social community, morphological (i.e., physical) variation, and a variety of other factors. Morphological variation, being by definition a fundamental influence on kinematics, holds potential as an explanation for interspeaker differences even in a seemingly fundamental law of motor control and behavior like Fitts' law. It is known that speakers vary widely in terms of a number of morphological characteristics, including vocal tract length (Vorperian et al., 2005, 2009) and relative proportions (Fitch & Giedd, 1999; Vorperian & Kent, 2007), as

well as hard palate and posterior pharyngeal wall shape (Lammert, Proctor, & Narayanan, 2013b), and many other parameters. There is growing evidence that differences in morphology of the speech apparatus all influence the production of specific speech sounds at the level of articulatory goals and kinematics (Dart, 1991; Brunner et al., 2009; Fuchs et al., 2008; Lammert, Proctor, & Narayanan, 2013a). A particularly intuitive example comes from indications that individuals vary in terms of their tongue size relative to the size of the entire speech apparatus (Lammert, Hagedorn, et al., 2013). It seems reasonable to expect that a smaller relative tongue size will result in longer articulatory distances travelled within the oral and pharyngeal cavities, on average, resulting in a wide range of values for ID. This wider range of ID might, in turn, cause the relationship between ID and MT to stand out against any noise in the data. The potential effects of morphology also points at specific hypotheses. For instance, it has already been discussed in the present work how low-back vowels appear to be the most difficult vowels to produce, and it was suggested that this may be the result of longer articulatory distances associated with producing them. It has been well-documented that males have a proportionally longer pharynx than females (Vorperian et al., 2011), which would amplify the distances required for the tongue to travel, causing further increases in ID associated with low-back vowels, and likely a wider range of ID overall. The potential connection between vocal tract morphology and Fitts' law for speech production merits further attention.

It should be noted that there are many sources of variability in the present analysis that may have had an impact on the correlation values, and may limit the generality of these results. One limitation relates to the accuracy of finding a video frame near the temporal center of a given phone, which is limited by the temporal resolution of rtMRI and the quality of forced phoneme alignment. Recent advances in rtMRI protocols may alleviate this limitation (Lingala, Sutton, et al., 2016; Lingala, Zhu, et al., 2016). If speaking rate is a factor, then the correlation values from Table 1 should, in turn, be correlated with speaking rate. Speaking rate was computed for all subjects by looking at the mean time

between adjacent syllable nuclei to get an estimate of syllable rate. Pearson's correlations were found between these values and the correlation values for Nucleus–Coda Consonant ($r = -0.64$, $p = 0.047$) and Onset Consonant–Nucleus ($r = -0.63$, $p = 0.049$). The fact that speaking rate is a factor indicates that the current data may have frame rates that are at the boundary of usefulness for the analysis done in this study. Higher frame rates would be preferable in future work. Additional variability may stem from non-Gaussian noise on pixel intensity values that rtMRI images often contain. Added variability in the data and analysis would have the clearest impact on the Onset-Onset and Coda-Coda task results, due to their much smaller number. Data are also limited to a midsagittal view of the speech articulators, meaning not all kinematic aspects are captured in the data.

Conclusion

This paper has presented an analysis of speech articulation from a large database of real-time magnetic resonance (rtMRI) data, in order to assess whether articulatory kinematics conform to Fitts' law. It appears that certain aspects of speech production do conform to Fitts' law, with the strength of that relationship varies across speaker and context-target type. The strongest such relationships are seen for VC context-target tasks, with CV tasks showing nearly as strong correlations. Also presented was a novel methodology for addressing the challenges inherent in performing Fitts-style analysis on rtMRI data of speech production, from defining the key quantities to extracting them from rtMRI data. Finally, a novel mathematical argument was presented for the expectation of Fitts' law in speech production, and why one expects to observe behavior consistent with the law on the basis of Task Dynamics and the VITE neural model of directed movement. Future work should focus on addressing the remaining methodological challenges. Among these challenges are higher frame rate data, and exploring additional definitions of the key relevant quantities.

Acknowledgements

This work is sponsored by the Assistant Secretary of Defense for Research & Engineering (ASD[R&E]) under Air Force contract #FA8721-05-C-0002. Funding also provided by the National Science Foundation (#1514544). The authors would like to thank Dr. Benjamin Parrell for sharing his insights on Task Dynamics and possible connections to neural dynamics.

References

Beamish, D., Bhatti, S. A., MacKenzie, I. S., & Wu, J. (2006). Fifty years later: a neurodynamic explanation of Fitts' law. *Journal of The Royal Society Interface*, *3*(10), 649–654.

Bresch, E., Nielsen, J., Nayak, K., & Narayanan, S. (2006). Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans. *The Journal of the Acoustical Society of America*, *120*(4), 1791–1794.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*(3-4), 155–180.

Browman, C. P., & Goldstein, L. (1995). Gestural syllable position effects in american english. *Producing speech: Contemporary issues*, 19–33.

Brunner, J., Fuchs, S., & Perrier, P. (2009). On the relationship between palate shape and articulatory behavior. *The Journal of the Acoustical Society of America*, *125*(6), 3936–3949.

Bullock, D., & Grossberg, S. (1988). Neural dynamics of planned arm movements: Emergent invariants and speed-accuracy properties during trajectory formation. *Psychological Review*, *95*(1), 49–90.

Byrd, D., Tobin, S., Bresch, E., & Narayanan, S. (2009). Timing effects of syllable structure and stress on nasals: a real-time mri examination. *Journal of Phonetics*, *37*(1), 97–110.

Card, S. K., English, W. K., & Burr, B. J. (1978). Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys for text selection on a CRT. *Ergonomics*, *21*(8), 601–613.

Carpenter, R. H. (1988). *Movements of the eyes (2nd rev.* Pion Limited.

Chin, S. B., & Pisoni, D. B. (1997). *Alcohol and speech.* Academic Press.

Crossman, E., & Goodeve, P. (1983). Feedback control of hand movement and Fitt's law. *Quarterly Journal of Experimental Psychology*, *35A*, 251–278.

Dart, S. N. (1991). *Articulatory and acoustic properties of apical and laminal articulations* (Vol. 79). UCLA.

Davids, K., Glazier, P., Araújo, D., & Bartlett, R. (2003). Movement systems as dynamical systems. *Sports medicine*, *33*(4), 245–260.

Drewes, H. (2013). *A lecture on fitts' law.* Retrieved from `http://www.cip.ifi.lmu.de/ drewes/science/fitts/`

Drury, C. G. (1975). Application of Fitts' Law to foot-pedal design. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *17*(4), 368–373.

Duarte, M., & Freitas, S. (2005). Speed-accuracy trade-off in voluntary postural movements. *Motor control*, *9*(2), 180–196.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, *106*(3), 1511–1522.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, *47*(6), 381-391.

Fitts, P. M., & Peterson, J. (1964). Information capacity of discrete motor responses. *Journal of experimental psychology*, *67*(2), 103.

Fitts, P. M., & Radford, B. K. (1966). Information capacity of discrete motor responses under different cognitive sets. *Journal of Experimental Psychology*, *71*(4), 475.

Fuchs, S., Winkler, R., & Perrier, P. (2008). Do speakers' vocal tract geometries shape their articulatory vowel space? In *8th international seminar on speech production, issp'08* (pp. 333–336).

Guenther, F. H., Hampson, M., & Johnson, D. (1998). A theoretical investigation of reference frames for the planning of speech movements. *Psychological review*, *105*(4), 611.

Hardcastle, W. (1976). *Physiology of speech production: An introduction for speech scientists*. London: Academic Press.

Houde, J. F., & Nagarajan, S. S. (2011). Speech production as state feedback control. *Frontiers in human neuroscience*, *5*, 82.

Kato, T., Lee, S., & Narayanan, S. (2009). An analysis of articulatory-acoustic data based on articulatory strokes. In *International conference on acoustics, speech & signal processing* (pp. 4493–4496).

Katsamanis, A., Black, M., Georgiou, P. G., Goldstein, L., & Narayanan, S. (2011). SailAlign: Robust long speech-text alignment. In *Proc. of workshop on new tools and methods for very-large scale phonetics research.*

Kleinow, J., Smith, A., & Ramig, L. O. (2001). Speech motor stability in ipdeffects of rate and loudness manipulations. *Journal of Speech, Language, and Hearing Research*, *44*(5), 1041–1051.

Krakow, R. (1999). Physiological organization of syllables: A review. *Journal of Phonetics*, *27*, 23–54.

Krause, J. C., & Braida, L. D. (2002). Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *The Journal of the Acoustical Society of America*, *112*(5), 2165–2172.

Kuwabara, H. (1996). Acoustic properties of phonemes in continuous speech for different speaking rate. In *Icslp 96* (Vol. 4).

Lammert, A. C., Goldstein, L., Ramanarayanan, V., & Narayanan, S. (2014). Gestural control in the English past-tense suffix: an articulatory study using real-time MRI. *Phonetica*, *71*(4), 229–248.

Lammert, A. C., Hagedorn, C., Proctor, M., Goldstein, L., & Narayanan, S. (2013). Interspeaker variability in relative tongue size and vowel production. *The Journal of the Acoustical Society of America*, *134*(5), 4205–4205.

Lammert, A. C., Proctor, M., & Narayanan, S. (2013a). Interspeaker variability in hard palate morphology and vowel production. *Journal of Speech, Language, and Hearing Research*, *56*(6), S1924–S1933.

Lammert, A. C., Proctor, M., & Narayanan, S. (2013b). Morphological variation in the adult hard palate and posterior pharyngeal wall. *Journal of Speech, Language, and Hearing Research*, *56*(2), 521–530.

Lammert, A. C., Proctor, M. I., Narayanan, S. S., et al. (2010). Data-driven analysis of realtime vocal tract MRI using correlated image regions. In *Interspeech* (pp. 1572–1575).

Lammert, A. C., Ramanarayanan, V., Proctor, M. I., Narayanan, S., et al. (2013). Vocal tract cross-distance estimation from real-time MRI using region-of-interest analysis. In *Interspeech* (pp. 959–962).

Lammert, A. C., Shadle, C. H., Narayanan, S. S., & Quatieri, T. F. (2016). Investigation of speed-accuracy tradeoffs in speech production using real-time magnetic resonance imaging. *Interspeech 2016*, 460–464.

Langolf, G. D., Chaffin, D. B., & Foulke, J. A. (1976). An investigation of fitts' law using a wide range of movement amplitudes. *Journal of Motor Behavior*, *8*(2), 113–128.

Lehiste, I. (1970). *Suprasegmentals*. MIT Press.

Lingala, S., Sutton, B., Miquel, M., & Nayak, K. (2016). Recommendations for real-time speech MRI. *Journal of Magnetic Resonance Imaging*, *43*(1), 28–44.

Lingala, S., Zhu, Y., Kim, Y., Toutios, A., Narayanan, S., & Nayak, K. (2016). A fast and flexible MRI system for the dynamic study of vocal tract shaping. *Magnetic Resonance in Medicine*.

Lofqvist, A., & Gracco, V. L. (1997). Lip and jaw kinematics in bilabial stop consonant production. *Journal of Speech, Language, and Hearing Research*, *40*(4), 877–893.

MacKenzie, I. S. (1992). Fitts' law as a research and design tool in human-computer interaction. *Human-computer interaction*, *7*(1), 91–139.

MacNeilage, P. (1972). *Speech physiology*. University of Texas.

Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A. C., Kim, J., Lee, S., ... others (2014). Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). *The Journal of the Acoustical Society of America*, *136*(3), 1307–1311.

Perrier, P., & Fuchs, S. (2008). Speed–curvature relations in speech production challenge the 1/3 power law. *Journal of Neurophysiology*, *100*(3), 1171–1183.

Ramanarayanan, V., Goldstein, L., Byrd, D., & Narayanan, S. S. (2013). An investigation of articulatory setting using real-time magnetic resonance imaging. *The Journal of the Acoustical Society of America*, *134*(1), 510–519.

Ramanarayanan, V., Lammert, A. C., Goldstein, L., & Narayanan, S. (2014). Are articulatory settings mechanically advantageous for speech motor control? *PloS one*, *9*(8), e104168.

Saltzman, E. L., & Kelso, J. (1987). Skilled actions: a task-dynamic approach. *Psychological review*, *94*(1), 84.

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological psychology*, *1*(4), 333–382.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J., Weninger, F., & Eyben, F. (2014). Medium-term speaker states: A review on intoxication, sleepiness and the first challenge. *Computer Speech & Language*, *28*(2), 346–374.

Shannon, C., & Weaver, W. (1949). *The mathematical theory of information*. University of Illinois Press.

Sibert, L. E., & Jacob, R. J. (2000). Evaluation of eye gaze interaction. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 281–288).

Templin, M. (1957). *Certain language skills in children; their development and interrelationships.* University of Minnesota Press.

Turvey, M. T. (1990). Coordination. *American psychologist*, *45*(8), 938.

Vorperian, H. K., & Kent, R. D. (2007). Vowel acoustic space development in children: A synthesis of acoustic and anatomic data. *Journal of Speech, Language, and Hearing Research*, *50*(6), 1510–1545.

Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., & Yandell, B. S. (2005). Development of vocal tract length during early childhood: A magnetic resonance imaging study. *The Journal of the Acoustical Society of America*, *117*(1), 338–350.

Vorperian, H. K., Wang, S., Chung, M. K., Schimek, E. M., Durtschi, R. B., Kent, R. D., ... Gentry, L. R. (2009). Anatomic development of the oral and pharyngeal portions of

the vocal tract: An imaging study a. *The Journal of the Acoustical Society of America*, *125*(3), 1666–1678.

Vorperian, H. K., Wang, S., Schimek, E. M., Durtschi, R. B., Kent, R. D., Gentry, L. R., & Chung, M. K. (2011). Developmental sexual dimorphism of the oral and pharyngeal portions of the vocal tract: An imaging study. *Journal of Speech, Language, and Hearing Research*, *54*(4), 995–1010.

Ware, C., & Mikaelian, H. H. (1987). An evaluation of an eye tracker as a device for computer input. In *Acm sigchi bulletin* (Vol. 17, pp. 183–188).

Welford, A. (1968). *Fundamentals of Skill*. Methuen.

Williamson, J. R., Quatieri, T. F., Helfer, B. S., Perricone, J., Ghosh, S. S., Ciccarelli, G., & Mehta, D. D. (2015). Segment-dependent dynamics in predicting Parkinson's disease. In *Sixteenth annual conference of the international speech communication association.*

Wrench, A. (1999). *The mocha-timit articulatory database.* Retrieved from `http://www.cstr.ed.ac.uk/artic/mocha.html`

Yunusova, Y., Weismer, G., Westbury, J. R., & Lindstrom, M. J. (2008). Articulatory movements during vowels in speakers with dysarthria and healthy controls. *Journal of Speech, Language, and Hearing Research*, *51*(3), 596–611.

	Onset - Nucleus	Nucleus - Coda	Onset - Onset	Coda - Coda	Coda - Onset
M1	r = -0.06 p=0.48 (n=141)	r = 0.21* p<0.05 (n=92)	r = -0.01 p=0.96 (n=29)	r = -0.08 p=0.72 (n=25)	r = 0.05 p=0.40 (n=266)
M2	**r = 0.49***** p<0.001 (n=183)	**r = 0.72***** p<0.001 (n=127)	r = -0.13 p=0.48 (n=33)	**r = 0.52***** p<0.001 (n=40)	**r = 0.37***** p<0.001 (n=384)
M3	r = 0.03 p=0.70 (n=187)	**r = 0.29***** p<0.001 (n=125)	r = -0.25 p=0.18 (n=32)	r = -0.28 p=0.07 (n=41)	r = 0.05 p=0.34 (n=411)
M4	r = 0.07 p=0.32 (n=184)	**r = 0.33***** p<0.001 (n=126)	r = -0.35* p<0.05 (n=33)	r = -0.30 p=0.05 (n=42)	**r = 0.15**** p<0.01 (n=413)
M5	r = -0.01 p=0.89 (n=147)	**r = 0.38***** p<0.001 (n=97)	r = -0.21 p=0.26 (n=30)	r = 0.18 p=0.38 (n=26)	r = 0.03 p=0.56 (n=312)
F1	**r = 0.36**** p<0.01 (n=183)	**r = 0.41***** p<0.001 (n=126)	r = 0.02 p=0.91 (n=32)	**r = 0.46**** p<0.01 (n=41)	**r = 0.24***** p<0.001 (n=389)
F2	**r = 0.30***** p<0.001 (n=182)	**r = 0.49***** p<0.001 (n=127)	**r = 0.40**** p<0.01 (n=33)	r = 0.38* p<0.05 (n=41)	**r = 0.17***** p<0.001 (n=388)
F3	r = -0.04 p=0.56 (n=182)	**r = 0.30***** p<0.001 (n=126)	r = 0.20 p=0.27 (n=32)	r = -0.33* p<0.05 (n=41)	r = 0.05 p=0.34 (n=413)
F4	r = -0.11 p=0.19 (n=153)	**r = 0.31**** p<0.01 (n=102)	r = -0.26 p=0.17 (n=30)	r = -0.18 p=0.37 (n=28)	r = -0.03 p=0.60 (n=291)
F5	r = 0.06 p=0.41 (n=184)	**r = 0.28**** p<0.01 (n=126)	r = -0.12 p=0.50 (n=32)	r = -0.04 p=0.80 (n=40)	r = 0.06 p=0.26 (n=407)

Table 1

Pearson's r (and p-values) between movement time (MT) and index of difficulty (ID) for all subjects, divided by syllable position-specific category. Correlation coefficients significant at the $\alpha = 0.05$*,* $\alpha = 0.01$ *and* $\alpha = 0.001$ *level are marked with* $*$*,* $**$ *and* $***$*, respectively.*

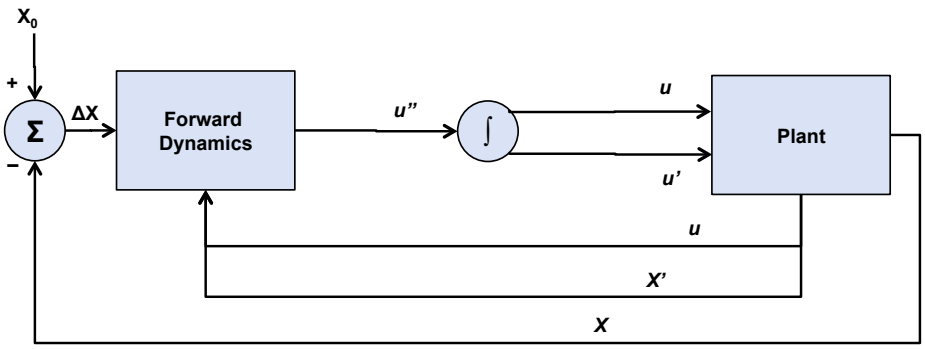


Figure 1. Schematic representation of the Task Dynamics framework. The variable X is the displacement of the controlled variable in task space and X_0 is the target. The Forward Dynamics component implements a second-order dynamical system, conforming to Equation 3, that transforms (via inverse kinematics) the error signal, ΔX, into the second derivative of the articulator-space variable u. The integrals $\dot{u}$ and u function as motor commands to the Plant, or speech production apparatus.

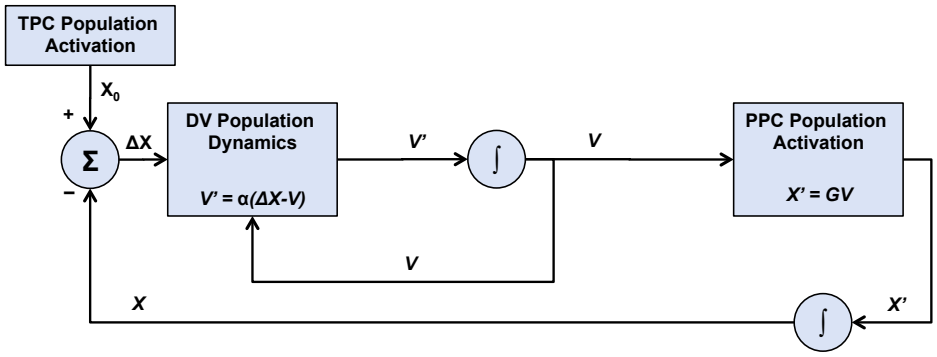


Figure 2. Schematic representation of the VITE neural model (Bullock & Grossberg, 1988). Note the many similarities of this structure to that of Task Dynamics in Figure 1. TPC is a representation of the target position, which produces a target position X_0*. The DV population compares the target to the system's current position, and computes the task-space dynamics of the network. The PPC population integrates the DV population activation into position information. The network dynamics have the form described in Equations 6 and 7.*

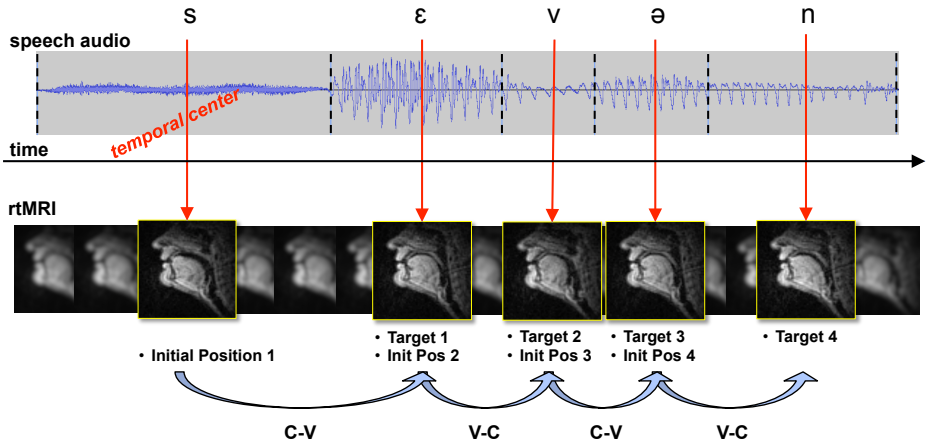
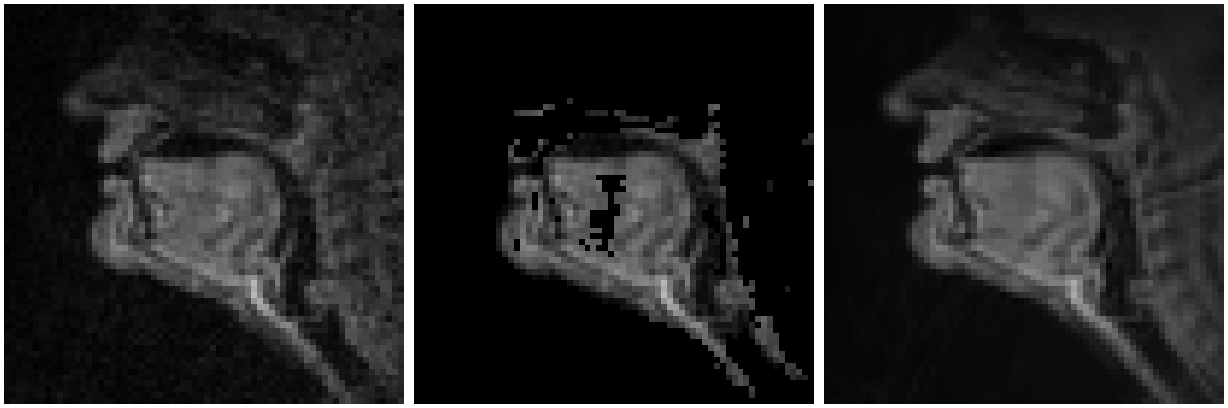


Figure 3. A key concept behind the methodology developed in the present work is that motor tasks in speech articulation can be viewed as a sequence of movements toward and away from target points in articulatory space. Those targets are assumed to be approached and approximated, but not necessarily reached, at the temporal center of each phone interval. The initial position for a given task is assumed to be the target immediately preceding the current one.



(a) (b) (c)

Figure 4. Images illustrating stages in the data pre-processing pipeline for a single vocal tract posture. Shown are (a) an image of a single posture, in its original form, (b) the same image with low-variance pixels masked out (c) the image again, reconstructed as an image, but using only the L PCA-generated features.

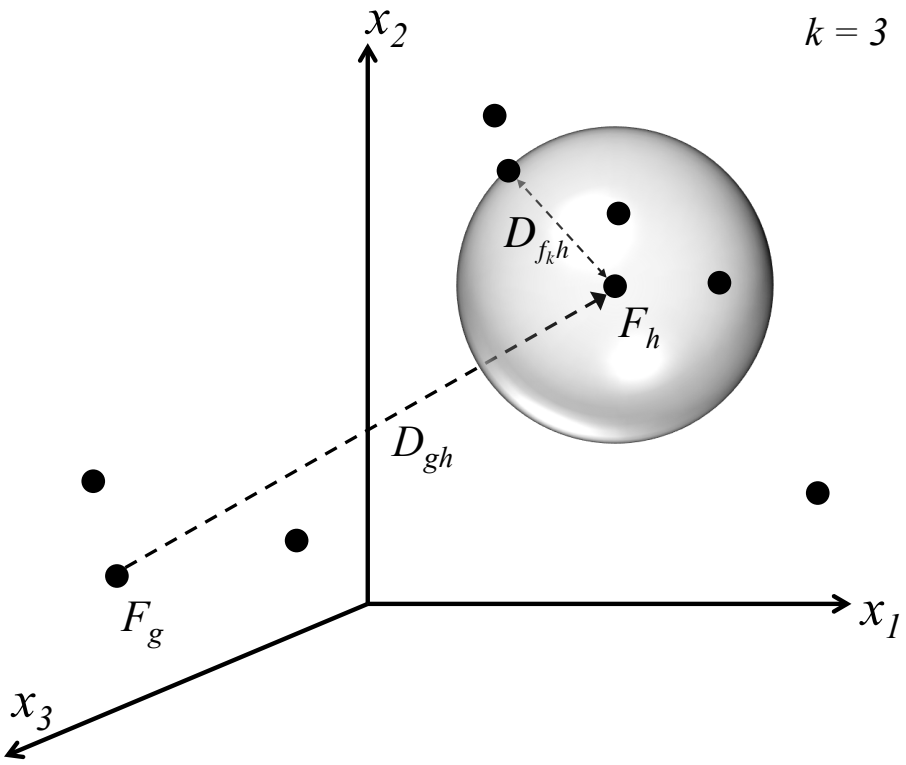


Figure 5. Illustration of the key relationships in calculating ID from articulatory data, with most variable names taken from the text. Target vectors are defined in the high-dimensional articulatory space, represented in the illustration by features x_1, x_2, x_3. In the analysis, this articulatory space is actually composed of L total features. The articulatory target vector F_g is the target of the previous movement, and represents the starting point of the current movement. The target of the current movement is F_h. The distance to the target is the Euclidean distance between these two vectors. The width around the target is calculated with respect to a hypersphere around the current target, which is used to estimate the density of other target vectors that are not the current one.

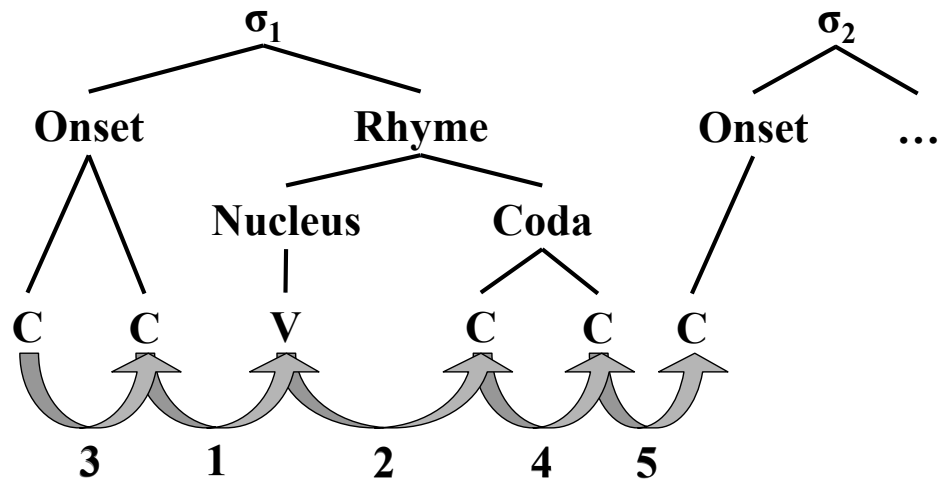
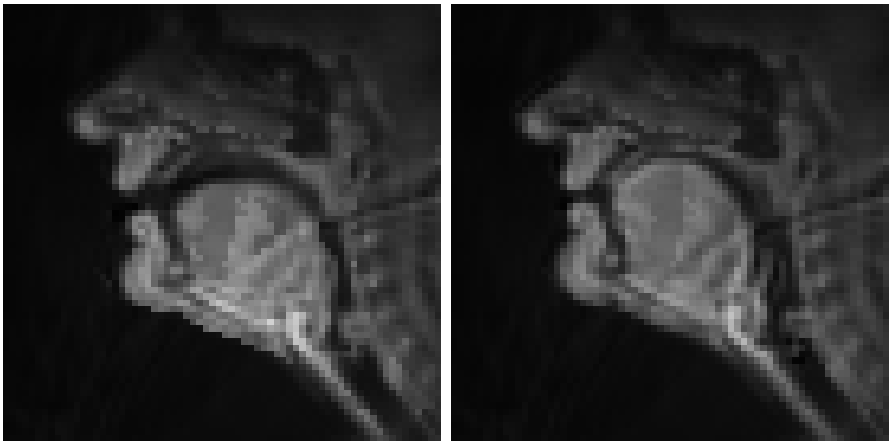
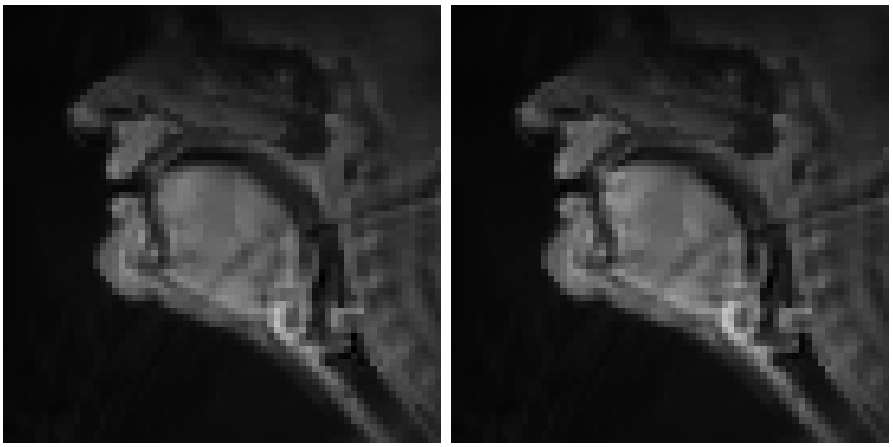


Figure 6. Illustration of the different syllable position-specific task categories used in the present analysis, shown on a traditional, generic syllable structure tree. Categories are numbered outward from the nucleus, and include tasks leading into and out of the nucleus (1 & 2), tasks between consonants in the onset and coda (3 & 4) and tasks leading from one syllable to the next (5).

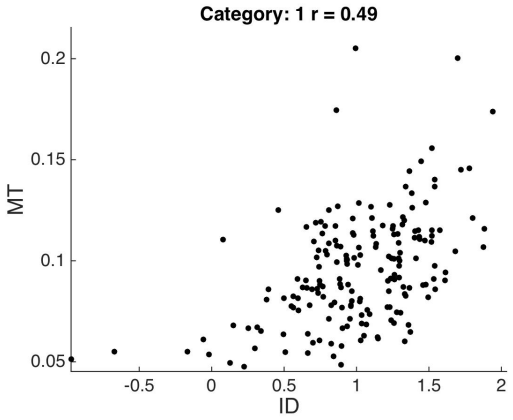


(a) *initial position:* ɑ (b) *target:* ʃ

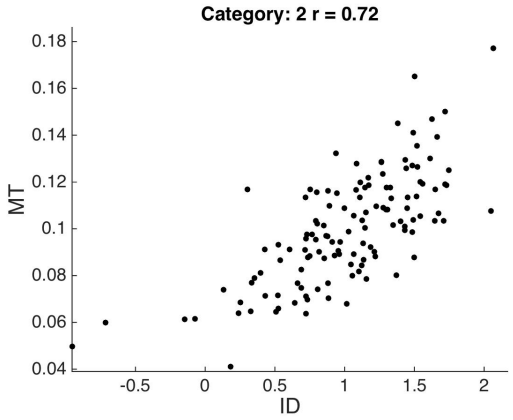


(c) *initial position:* i (d) *target:* d

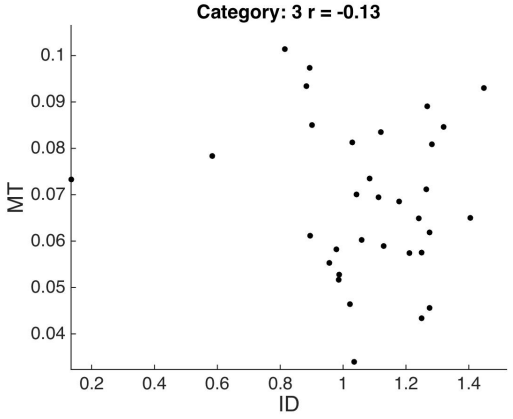
Figure 7. Example high- and low-ID tasks for subject M2. The top row, (a)-(b), represent one of the highest ID tasks, while the bottom row (c)-(d) represents one of the lowest. Images were reconstructed from the L articulatory features in Z (see text).



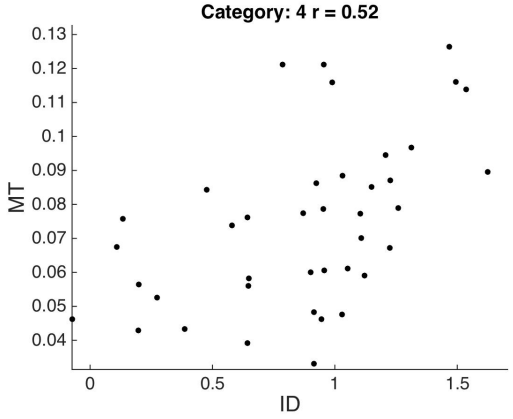
(a) Category 1: Onset-Nucleus Task



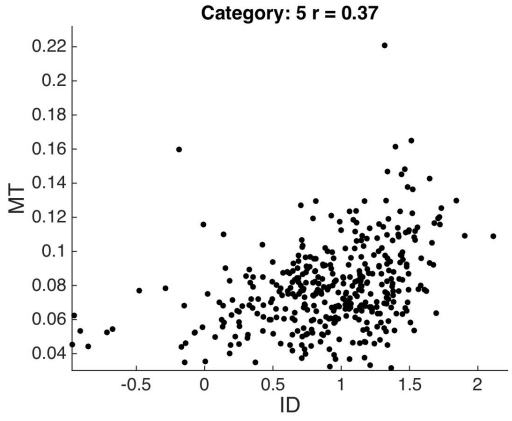
(b) Category 2: Nucleus-Coda Task



(c) Category 3: Onset-Onset Task



(d) Category 4: Coda-Coda Task



(e) Category 5: Coda-Onset Task

Figure 8. Movement time (MT) vs. index of difficulty (ID) for subject M2. All context-target tasks are shown, divided by syllable position-based category (see text for details concerning categories).